A REPORT

ON

RULE GENERATION USING ASSOCIATION RULE MINING

BY

ANIKETH JANARDHAN REDDY	2014A7PS096H
KARTIK SETHI	2014A7PS130H
MONICA ADUSUMILLI	2014A7PS005H
VINAY DATTA	2014A7PS509H

Under the supervision of

Prof. N. L. Bhanu Murthy

Department of Computer Science and Information Systems

Submitted as a part of the academic project for the course

CS F415 Data Mining

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

LANGUAGE USED

C++

DATA PRE PROCESSING

The data has 17 attributes each of which can have the value of yes or no. Each of the yes or no is taken as an item. For example, className=democrat is an item and className=Republican is an item. Hence, a total of 34 items are made to make the data suitable for rule generation.

COMPILATION STEPS

$ g++ generatefrequentitemsets_ap.cpp

$ ./a.out

Input: vote.arff **Output:** frequent_itemsets.txt

The input file is vote.arff which contains the vote data (input file given) and it outputs the number of frequent item sets along with the frequent item sets in the file frequent_itemsets.txt for the support mentioned in the file generatefrequentitemsets_ap.cpp.

$ g++ generaterules.cpp

$ ./a.out

Input: frequent_itemsets.txt and valtoattr.txt **Output:** rules.txt

The input file is the frequent_itemsets.txt and also it takes input from the valtoattr.txt which has the attribute number and the actual string name of the attribute and the output file is the rules.txt which generates the association rules for the confidence mentioned in the file generaterules.cpp.

SUPPORT AND CONFIDENCE VALUES, NUMBER OF RULES GENERATED

SUPPORT VALUE	CONFIDENCE VALUE	FREQUENT ITEMSETS	ASSOCIATION RULES
0.35	0.8	321	1502
0.35	0.9	321	735
0.38	0.7	176	862
0.38	0.8	176	630
0.38	0.9	176	306
0.4	0.8	118	383
0.4	0.9	118	193
0.5	0.8	17	14
0.5	0.9	17	6
0.6	0.7	4	0

At a support of 0.5 and a confidence of 0.8, 17 frequent item sets were found and 14 association rules were mined and at a support of 0.45 and a confidence threshold of 0.95, 44 frequent item sets were found and 11 rules are generated. These were some of the interesting rules generated by experimenting with different values of support and confidence.

It is also observed that very low support and very high support result in too many or too less frequent item sets. Also very low confidence or very high confidence is equally bad. Hence it is important to have an optimum values of support and confidence. Low values of confidence yield more number of association rules than the frequent item sets which again is sub optimal. A confidence of 0.8-0.95 and a support of 0.45-0.5 is found as the optimal one in this case.

It is also important to optimize both support and confidence values. A low value of support with optimal confidence or low confidence with optimal support hasn't given the best of the rules.

With a low support value of 0.35, and an optimal confidence value of 0.8, 321 frequent item sets and 1502 were generated. Such rules give us way too many insights than required and might not be optimal.

ASSOCIATION RULES

Some of the interesting rules generated at support of 0.45 and confidence of 0.95 are below.

Frequent Item sets = 44 Rules=11

The rules are as follows:

Confidence = 1 Rule: adoption-of-the-budget-resolution=y, physician-fee-freeze=n -> Class=republican

Confidence = 1 Rule: adoption-of-the-budget-resolution=y, physician-fee-freeze=n, aid-to-nicaraguan-contras=y -> Class=republican

Confidence = 0.975369 Rule: adoption-of-the-budget-resolution=y, aid-to-nicaraguan-contras=y, Class=republican -> physician-fee-freeze=n

Confidence = 0.991903 Rule: physician-fee-freeze=n -> Class=republican

Confidence = 0.995261 Rule: physician-fee-freeze=n, aid-to-nicaraguan-contras=y -> Class=republican

Confidence = 0.99505 Rule: physician-fee-freeze=n, education-spending=n -> Class=republican

Confidence = 0.980769 Rule: el-salvador-aid=n -> aid-to-nicaraguan-contras=y

Confidence = 0.961538 Rule: el-salvador-aid=n -> Class=republican

Confidence = 0.965686 Rule: el-salvador-aid=n, aid-to-nicaraguan-contras=y -> Class=republican

Confidence = 0.985 Rule: el-salvador-aid=n, Class=republican -> aid-to-nicaraguan-contras=y

Confidence = 0.963303 Rule: aid-to-nicaraguan-contras=y, Class=republican -> physician-fee-freeze=n